**Predictive Analysis for iPhone Purchases**

**Introduction:**

In recent years, the mobile phone industry has witnessed unprecedented growth and competition, marked by rapid technological advancements and changing consumer preferences. The smartphone market, particularly the segment dominated by Apple's iPhones, remains dynamic and highly responsive to shifts in consumer behaviour. As the industry landscape evolves, businesses face the challenge of tailoring their strategies to align with customer expectations, maximize market share, and optimize resource utilization.

**Objective:**

The objective of the project is to develop a predictive model that can determine whether an individual is likely to purchase an iPhone based on their demographic and financial information.

**Methodology:**

* **Data Collection:** The data used to build the model was provided in the csv format. It has gender, age, salary and whether the said person purchased the phone or not in 1 and 0 format, where 1 stands for yes and 0 for no.
* **Data Pre-processing:** Data pre-processing is a crucial step to ensure the quality and suitability of the dataset for training machine learning models.
* **Feature Selection:** Feature selection is a critical step to identify the most relevant variables that contribute to the predictive power of the model.
* **Model Selection:** In the model selection section, provide a detailed overview of the machine learning algorithms chosen for the predictive analysis. Explain the rationale behind the selection of each algorithm and discuss how they align with the project objectives.
* **Model Training:** In the model training section, the processed data is fit to train the selected model so that it is able to predict the future entered data.
* **Model Evaluation:** In the model evaluation section, the performance of the trained machine learning models is assessed to select the best suited model for deployment.

**Algorithm Used:**

* **KNeighborsClassifier** - KNeighborsClassifier is a classification algorithm based on the K-Nearest Neighbors (KNN) approach. It is part of the scikit-learn library in Python and is used for solving classification problems. KNN is a simple and intuitive algorithm that classifies a new data point based on the majority class of its k-nearest neighbors in the feature space.
* **Decision Tree** - A decision tree is a supervised machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the dataset into subsets based on the values of different features.

The objective is to create a tree-like structure of decisions, where each node represents a decision based on a particular feature, and each leaf node represents the predicted outcome.

* **Random Forest** - A Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and

outputting the class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees.

It is a powerful and versatile machine learning algorithm known for its high accuracy and robustness.

* **LogisticRegression** - Logistic Regression is a statistical method used for binary classification. Despite its name, it is a classification algorithm rather than a regression algorithm. It's commonly used in machine learning for predicting the probability that an instance belongs to a particular category.

**Solution Architecture:**

* **Data Collection and Storage:** The data for the project was provided to us in a csv format. Using ‘pd.read\_csv()’ the data was used to read in the jupyter notebook for further processing.
* **Data Pre-processing:** In the data pre-processing stage, it is important to look for the missing or null values as well as duplicated data which might make the model less accurate. In this case the dataset was more cleaned with no null values.
* **Feature Selection:** The data had gender, age, salary and purchase iphone column with 1 and 0 values where 1 means yes and 0 means no. In this step it is necessary to select the features or variables properly to train and test the model for accurate prediction. The age and salary column are selected as x independent variable and purchase iphone column is selected as y dependent variable which will be predicted. Upon endocing gender as 0 and 1 values and using it along with age and salary as x variable gave less accurate result. So, it was discarded and only age and salary is used as x variable which is giving more accurate results.
* **Model Selection:** Model selection is based on the type of dataset or taking the problem in mind. Here, we need to predict whether a person will make a purchase or not in the 1 and 0 form where 1 means yes and 0 means no. So, it is a classification based problem. So, KNeighborsClassifier, Decision Tree and Random Forest algorithms are used.
* **Model Training:** Then the models are fitted on the dataset and trained to predict for the test dataset.
* **Model Evaluation:** Next step requires for the evaluation of the model for its accuracy using different metrics such as r2\_score, accuracy\_score, precision\_score, etc. Here, on evaluation the KNeighborsClassifier gave an accuracy of 87.5% , Logistic Regression model gave an accuracy of 66.25%, Decision Tree model gave an accuracy of 90% and Random Forest model gave an accuracy of o 87.5%.
* **Best Model Selection:** Then the model with highest accuracy is selected and saved for future use.